J. Med. Chem. **1997**, *40*, 3926−3936

New Methodology for Profiling Combinatorial Libraries and Screening Sets: Cleaning Up the Design Process with HARPick

Andrew C. Good and Richard A. Lewis*

Rhône-Poulenc Rorer, Rainham Road South, Dagenham, Essex, RM10 7XS, United Kingdom

Received June 18, 1997[⊗]

Combinatorial chemistry is a tool of increasing importance in the field of ligand design, as it can yield huge increases in the number of compounds available for screening. Unfortunately, it is often the case that the number of molecules which could theoretically be constructed greatly exceeds potential synthesis and screening capacity. For this new technology to be fully exploited, it will become vital to design libraries with reference to the properties of compounds already in existence, if the added value of each new molecular collection is truly to be maximized. Similarly, if we are to take full advantage of the potential of combinatorial chemistry in lead optimization, it is important that our library design paradigms are flexible, with diversity scoring functions that can be modified to suit particular projects. Here these challenges are addressed through the introduction of a novel computer-aided library design tool known as HARPick (heuristic algorithm for reagent picking). The program is accessible to the bench chemist, and incorporates several significant advances over currently available approaches. These include product-based diversity calculations that can be constrained at the reagent level; diversity measures constructed from multiple descriptors; improved pharmacophore key information and full pharmacophore profiling of entire molecular databases. The potential of these improvements to aid in diversity profiling is illustrated through comparison with established methodology, and possible further enhancements are discussed.

Introduction

With the ever increasing role of combinatorial synthesis in medicinal chemistry, the requirement to create computational tools that aid in library design has become acute.[1] To meet this need, a number of groups have developed new methodology for diversity measurement and compound selection.[2−7] In this paper, we first highlight some concerns with existing methodology and then describe our efforts to overcome them using the HARPick (heuristic algorithm for reagent picking) program. The resulting advances are illustrated and discussed critically, and finally, potential areas for further improvement are presented.

Combinatorial Libraries

There are many different strategies for performing combinatorial chemistry,[1] and the most appropriate technology to use is dependent on the resources and goals of the project. In its broadest sense, combinatorial chemistry can be defined as the process of making all possible combinations of appropriate reagents using a given reaction. Using this definition, it is easy to see how the number of possible reaction products will greatly exceed the resources of all but the most profligate organization. If we consider the example of an amide condensation, choosing all the available reagents from a commercial catalogue could lead to the selection over 3000 amine synthons and 3000 acid synthons. Combining these two reagent sets would result in a total of 9 million products. (Throughout this paper, the starting point of any discussion will be a virtual combinatorial library in which all possible products have been enumerated in computers, with the goal of designing a much smaller library for actual synthesis.) Assuming a screening cost of $0.10 per compound, the cost of screening the library would approach $1 million, ignoring the need to store around 10 000 96-well plates, and the need to dispose of large quantities of waste (possibly radioactive) that would be generated by the screening. Such constraints demand that we adopt a strategy of designing a "rational" subset and begs the question of what is meant by the term "rational".

Measurement of Molecular Diversity. Many descriptors are available to describe molecular diversity in terms relevant to drug−receptor interactions.[2] These measures include reagent shape on a 3D lattice,[3] 2D/3D fingerprints,[4] and pharmacophore keys.[4−6] There have been a number of papers that have attempted to assess descriptor quality for diversity profiling.[4,7] Descriptors were ranked by their ability to discriminate active and inactive compounds within a number of medicinal chemistry project data sets. In these studies, it was suggested that 2D fingerprints and simple shape descriptors make better descriptors than other alternatives such as 3D pharmacophore fingerprints. From our own perspective, such assertions regarding descriptor quality are rather sweeping. 2D substructure searches are used routinely to extract analogues from databases.[8] Similarly, measurement of shape variation provides one of the staple descriptors of 3D-QSAR calculations.[9] A capacity to distinguish active from inactive analogues from a single biological screen at a nanomolar level, is hardly proof of an ability to discriminate between heterogeneous activity classes. Within a single activity class, differences as small as a methyl group can have significant effects on activity. The structural differences that exist between different receptors will tend to be much larger, however. Thus, to some extent, the results of such studies could have been predicted. Perhaps the best lesson to be drawn from these descriptor compari-

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1997.

S0022-2623(97)00403-2 CCC: $14.00

sons is that 2D and simple shape descriptors may be well suited to the design of lead optimization libraries.

Pharmacophore Descriptors. A pharmacophore, defined here as the critical geometric arrangement of molecular fragments required for binding,[10] provides an efficient descriptor for primary ligand−receptor interactions, defining a necessary but not sufficient condition for biological activity. When we reviewed the literature on searching for novel leads within 3D databases, the general consensus suggested that pharmacophores are the descriptors of choice.[11] Results of 3D flexible searching within databases of known compounds have proven this in a practical sense.[12] To us, it therefore seems reasonable to employ such descriptors for molecular diversity calculations, as libraries providing good coverage of accessible pharmacophore space should also prove a good source of new leads. Another major advantage of pharmacophores is the fact that they are a whole molecule descriptor, including within them the concept of conformational flexibility.[11]

Product Diversity and Reagent Diversity. The use of pharmacophores implies that diversity is measured on the products of the combinatorial reaction, rather than the reagents alone. This is a more computationally expensive choice, simply because of the numbers involved. There are good reasons for making this choice. Reagent-based descriptors and diversity assessments are based on the assumption of the properties of fragments being additive and independent when assessing diversity for each molecule in a combinatorial library.[3] This will not be the case for pharmacophore-based (or most other 3D) functions. In addition, it has been shown that, when profiling data sets employing descriptors derived from library products, the resulting compound selections cover diversity space more efficiently than comparable calculations utilizing reagents.[13] Third, care must be taken with reagent-based functions to make sure that they are suitable for interlibrary comparisons: this is the same argument as for the choice between clustering and partitioning data. In contrast, the pharmacophore descriptor is ideal for interlibrary comparisons. Given the relative speed of the calculations, our practical experience is that library design is not the rate-limiting step in combinatorial synthesis, so that there is a definite cost benefit to performing product-based calculations of the type that we describe below. For all of these reasons, since we are concerning ourselves primarily with such general screening libraries in this paper, we have chosen to use pharmacophores as our primary descriptor and have developed product-oriented methods, despite the extra computational cost.

Division of Descriptor Space. Two basic procedures are currently in use for dividing descriptor space. These involve the application of (i) clustering techniques and (ii) cell-based partitioning approaches for compound selection. Clustering methodology can be defined as the division of a group of objects into clusters with high intracluster similarity and intercluster dissimilarity. Such techniques have been utilized for many years for generating diverse compound sets for screening.[2−4,8] Partitioning involves the subdivision of property space into a number of regions (bins). Partition-based profiling is then generally defined as the selection of an object subset for maximal coverage of these property bins. This approach is already being applied extensively in compound selection calculations.[2,5,6,14,15]

Clustering has the advantage of cleanly dividing up data sets which distribute themselves discontinuously in property space. For general library design, however, we anticipate the saturation of property space, and hence discontinuity should not present a major problem. Partitioning techniques have the benefit of providing a convenient common frame of reference in property space, making comparison between different libraries a simple process. Another advantage of cell-based methodology is that calculation times tend to scale linearly with the number of molecules being processed, making the partitioning paradigm more suitable (faster) for large data sets. A problem with clustering specific to the descriptors used with these studies is that it is difficult to employ a pharmacophore fingerprint for clustering calculations (a potential problem encountered in one of the descriptor comparison studies[4]). This is because the fingerprint is (i) too sparse on a per molecule basis, making the similarity measure very discontinuous and (ii) too sensitive—the number of pharmacophores present in a given molecule varies as approximately the cube of the number of pharmacophore centers. Small molecular differences can thus potentially lead to large differences in fingerprint.

As we wished to employ pharmacophore descriptors over large data sets and include the ability to undertake interlibrary comparison, we chose to use a partition-based approach.

The Chem-Diverse Approach

A number of studies have been undertaken into the efficacy of pharmacophore triplets as molecular descriptors.[16,17] Recently a commercial program, Chem-Diverse,[5,6] has been developed to exploit pharmacophore triplet information in diversity profiling. Chem-Diverse provides a variety of useful tools for diversity assessment by pharmacophore and is becoming an industry standard for this aspect of combinatorial chemistry. The Chem-Diverse protocol for molecular diversity is based on trying to obtain the maximum coverage of pharmacophore space by potential combinatorial chemistry products (Figure 1).

While the methodology employed fits in with our general requirements of a partition-based approach for assessing pharmacophore diversity, the current version Chem-Diverse suffers from a number of technical drawbacks which need to be addressed. These are discussed below.

Library Profiling and Compound Selection Using Chem-Diverse. A central part of the Chem-Diverse compound selection procedure requires on-the-fly conformational analysis of all potential library products (see Figure 1). Any pharmacophores found are added to a single pharmacophore key, which describes the ensemble of selected molecules. Compounds are only selected if the set of pharmacophores they express overlaps with the ensemble key by less than a user-defined amount, that is, if the molecule contains a significant number of previously unseen pharmacophores. As a consequence, the results of such searches are dependent on the order in which the molecules are extracted from the database (analogous to the single-pass clustering algorithm[18]).

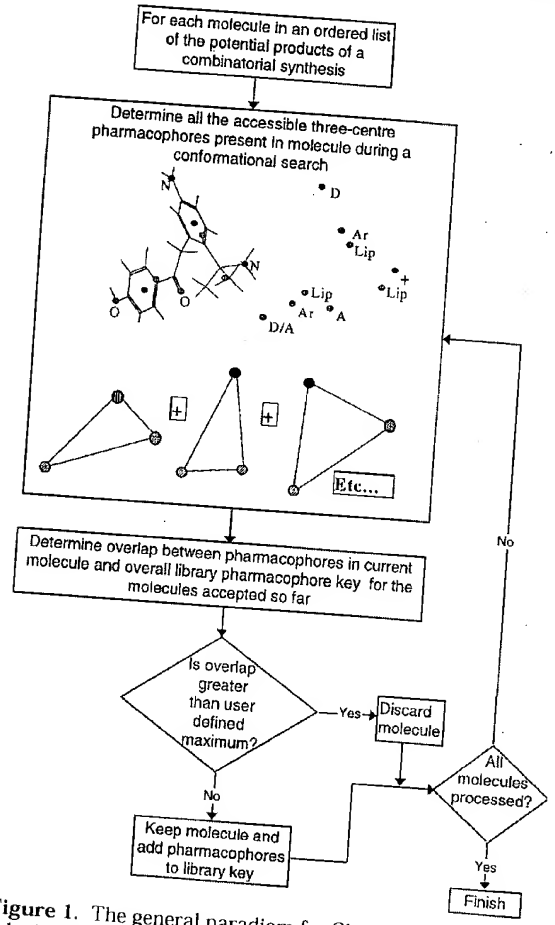

Figure 1. The general paradigm for Chem-Diverse profiling calculations. Key to pharmacophore centre types used here: A = hydrogen bond acceptor, D = hydrogen bond donor, Ar = aromatic ring centroid, Lip = lipophilic centroid (center of group of atoms with ~0 charge), + = charged positive.

There is no facility within Chem-Diverse to calculate diversity while constraining the number of reagents present in the selected products. This is crucial if control is to be exerted over the number of reagents required for any particular library synthesis. Instead, Chem-Diverse chooses molecules purely based on what it considers to be the most diverse set of products ("cherry-picking"), with no explicit reference to the constituent reagents. As a result, Chem-Diverse will often make an combinatorially inefficient selection of products. By inefficient we mean that, for example, when selecting 100 products from a two component combinatorial library (for example the amide condensation discussed earlier), rather than choose an 100% efficient 10 × 10 reagent set, Chem-Diverse will instead choose compounds comprising a larger reagent subset, say 30 × 20. Using such a selection would be both more costly and more complicated to program up on the synthesis robot and is thus termed inefficient. To achieve an efficient subset of products from a virtual combinatorial library using Chem-Diverse, selections must be analyzed by the user to determine the most frequently occurring reagents and repeated until eventually the final reagent choice is made.

Measurement of Diversity within Chem-Diverse. Within Chem-Diverse, modification of the search criteria to include additional molecular properties such as shape is not currently feasible. This is because the diversity function employed by Chem-Diverse cannot include any nonpharmacophoric properties. It is possible to remove unwanted reagents by assigning upper and lower bounds for given properties in their resulting products. Limitations in compound selection as described above makes such an approach rather risky, however, since it is not possible to devise an objective method of reagent removal which is entirely divorced from the product descriptors. This is important, because not all products created by a given reagent will necessarily be undesirable. Indeed, it may only be a few that are poor, with the remaining products adding much to the diversity of the library.

Limitations in Chem-Diverse Pharmacophore Keys. The Chem-Diverse pharmacophore keys are potentially extremely useful tools for directing library design. However, they are limited in their current incarnation, as the library key only registers whether or not a particular pharmacophore exists in the selected molecular ensemble, not how many times it is found. [The creation of a nonbinary key within Chem-Diverse has been developed since this work began. However, the current implementation (Oct96 version) still does not exploit the nonbinary pharmacophore data to constrain the construction of new libraries.] This makes the key prone to saturation, even when artificially small distance bins are applied, as is the case with the default bin settings in Chem-Diverse.

Methodology

To meet our requirements for diversity profiling, the HARPick program was developed. The basic outline of HARPick is illustrated in Figure 2. A number of features have been incorporated in the software to overcome many of the problems associated with Chem-Diverse. These are listed below.

Introduction of a Stochastic Optimization Algorithm. To remove the order dependence (product selection is dependent on the order in which the products are processed) of Chem-Diverse calculations and allow reagent selection direct from product diversity, an alternative technique of compound selection was required. We chose simulated annealing as our method, as it has a proven track record for locating good (and hopefully near global) minima on a complicated energy surface (in this case, "energy" is defined as the diversity function score), and could easily be incorporated into our diversity profiling paradigm. Our implementation is based on a standard simulated annealing algorithm,[19] employing fixed-length Markov chains and dynamic cooling.[20] Essentially, all changes in reagent selection which result in a reduction in the energy function (ΔE) are accepted, while changes producing a positive ΔE are accepted with a probability of $\exp(-\Delta E/T)$, where T is the annealing control temperature. In addition, a simple minimizer was also included. The minimization procedure only accepts reductions in ΔE and terminates after failing to find a new minimum for a user-defined number of Markov chains.

A number of studies have been undertaken which apply stochastic optimization techniques to the problem of diversity profiling. In general, optimization algorithms have been used as a substitute for deterministic techniques such as clustering.[21,22] Some attempts have

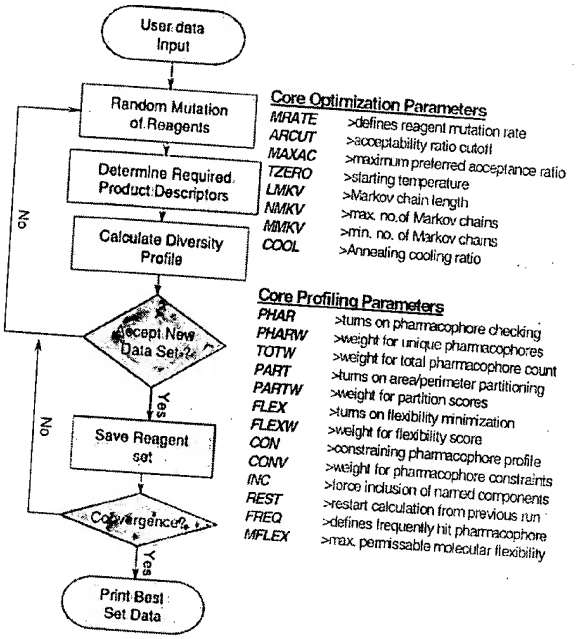

Figure 2. Structure of the HARPick program. The basic procedure is centred around a simulated annealing function which is used to make random selections of reagents from each of the components of combinatorial synthesis. Some of the core controls of the input file are shown. Particular features of the program include the following: (i) The separation of compound selection from the optimization function, allowing selection in reagent space and diversity calculation in product space. (ii) The facility to force the inclusion of named reagents from each component in the calculation. This allows a user defined bias to the selected data set including, for example, certain reagents deemed essential by the chemists. (iii) The capability to weight each of the functions used in the diversity calculation to suit user requirements. See the methodology section for more details.

been made to postprocess data from resulting diversity profile calculations, to determine frequently occurring reagents.[23] There have also been studies which highlight the theoretical potential of decoupling the reagent selection criterion from the diversity function calculation.[24] We have attempted to take the application of such methodology a stage further. As has been suggested, the introduction of a stochastic optimization technique makes the separation of object selection from scoring function calculation simple. This has two primary advantages.

(i) Reagent selection can now be divorced from the diversity evaluation. It is then possible to make selections in reagent space, while diversity is calculated in product space. This allows the user direct control over the number of reagents selected from each component pool of a combinatorial synthesis, rather than relying on the Chem-Diverse "cherry-picking" approach described above. This feature has been implemented in HARPick (Figure 2).

(ii) It is simple to introduce flexible scoring functions incorporating many diverse properties. Therefore, as well as including pharmacophores as our primary descriptor, it is also possible to add additional secondary descriptors. The real advantage of this is that such properties do not necessarily need to be made optimally diverse. Rather, they may be moderating descriptors designed to ensure sensible compound selection. A number of such measures have been applied and are described below.

Changes to Pharmacophore Profiling. For this work the basic Chem-Diverse pharmacophore descriptions are employed, using customized pharmacophore center, pharmacophore distance, and conformational search parameters. Each pharmacophore is specified by three interaction centers involving seven center types: (i) hydrogen bond donor, (ii) hydrogen bond acceptor, (iii) hydrogen bond donor and acceptor, (iv) aromatic, (v) hydrophobe, (vi) acidic, (vii) basic, leading to a total of 84 combinations of three centers. Each triangle edge distance is separated into 17 bins, leading to a total of 184 884 geometrically accessible pharmacophores. The number of distance bins used in the key creation has been adjusted to 17 instead of the 31 used in the default version of Chem-Diverse. This increased coarseness is felt to be justified when the large rotational increments of the Chem-Diverse conformational search procedures are considered. The 17 bins have been tailored to approximate the 20% tolerance determined experimentally for 3D databse searches involving rule-based conformational analyses.[25]

If we are to implement a stochastic optimization algorithm in reagent selection, pharmacophore keys for each molecule need to be stored on-line and accessed as and when required. It is not possible to use standard Chem-Diverse keys to store such data, since each key requires around 30 Kbytes of hard disk space. To overcome this, a Chem-X[5] PCL (program control language) script was written which profiles each data set molecule, extracts the Chem-Diverse key, decodes it, and writes out the individual pharmacophores found for the structure. Since the key for an individual structure is sparsely populated, a large saving in space can be made with such an approach, since the absent pharmacophores, which make up the bulk of the disk space, are ignored. Each pharmacophore written out requires 4 bytes of space; 1 for each of the three distances plus 1 for the pharmacophore type. Note that for each molecule, no single pharmacophore triplet can be hit more than once. This is of importance because it prevents particularly promiscuous molecules from skewing the pharmacophore distributions. Many, if not most, pharmacophores present in a molecule are small relative to the largest pharmacophore in the structure and are thus unlikely to explain the binding of that molecule to a particular receptor. It would therefore be useful if a technique were employed to remove these "insignificant" pharmacophores. Chem-Diverse provides a method which divides the area of any particular pharmacophore triangle by the number of heavy atoms in the molecule. Any triangles falling below a user-defined ratio for this value are removed from the calculation. The problem with such an approach is that the relationship between heavy atom count and pharmacophore triangle area is purely empirical. As a consequence, if the required ratio is set high to remove a large number of such "insignificant" pharmacophores, some molecules can have all their pharmacophores deleted. This happens when the largest pharmacophore area present in the structure is small relative to the number of heavy atoms. To get around this problem, we have implemented an alternative, self-consistent method for pharmacophore removal. The technique allows the user to set the minimum ratio

required for pharmacophore perimeter, relative to the largest perimeter found for all pharmacophores in the current structure. Since all calculations are carried out with respect to the internal pharmacophore geometries of each individual molecule, no structure can lose all its pharmacophores, only the ones defined as small.

Implementation of a Customizable Diversity Profiling Function. To fully exploit the application of a simulated annealing paradigm, a variety of useful functions have been incorporated into the diversity evaluation routine:

(i) The primary function, *Unique*, included within HARPick is equivalent to the scoring parameter employed by Chem-Diverse. That is, the *Unique* function keeps a count of the number of pharmacophore bins occupied in the selected set of library products. We are also employing a nonbinary description of pharmacophore space, which means that not only do we know which pharmacophores are hit, but also how many times. The *Unique* function thus corresponds to the number of non-zero variables in our pharmacophore integer array.

(ii) We have incorporated a partition function calculation to ensure the even distribution of three molecular properties which provide a crude measure of shape. (Note that we do not consider these descriptors as necessarily the best meaures of shape. They are, however, simple to calculate and do describe different aspects of molecular size and geometry. They also illustrate the ease with which diversity functions can be modified when using a stochastic optimization method for diversity analysis.) These are (a) number of heavy atoms (*ha*), (b) largest triangle perimeter present for all pharmacophores found (*pp*), (c) largest triangle area present for all pharmacophores found (*pa*). The minimum and maximum values for each property are determined for the entire library of potential products. The resulting property range is then divided into equal partitions between these bounds. During the diversity calculation, each selected molecule is assigned to a partition according to its property value. The number of molecules in each partition is then compared with the number expected for a perfectly flat distribution. The occupancy function is a maximum when each bin is equally occupied, in the hope of forcing the molecules in the generated product subset to have an even distribution of shapes or other properties, while still maximizing pharmacophore diversity.

$$\text{Partscore} = \frac{\max_o - \sum_{j=1}^{p}\sqrt{(\bar{o} - o_j)^2}}{\max_o} \quad (1)$$

where Partscore = partition score, $\max_o$ = maximum possible mean absolute deviation (when all molecules occupy a single partition), $\bar{o}$ = mean molecule occupation across all partitions p, o_j = number of molecules occupying partition j.

(iii) To allow control over molecular flexibility, a function incorporating the number of calculable conformations for each molecule (as defined by the conformational search criterion used in the Chem-Diverse profiling PCL script) has been included.

$$\text{Flex} = \frac{\sum_{i=1}^{n} f_i}{n} \quad (2)$$

where Flex = flexibility score, f_i = the number of calculable conformations for molecule i, n = number of selected molecules.

(iv) If we are to fully exploit the pharmacophoric data available, it is crucial that we move beyond binary key descriptions of libraries. As described above, the first step of our procedure is to compute and extract the pharmacophores for all (potential) products and store the results using Chem-Diverse. It is then a simple task to sum the resulting molecular profiles and produce an overall description of library pharmacophore coverage. This descriptor may then be employed as a constraint to optimize pharmacophore spread in any potential new libraries. To this end, a constraining function has been added to the profiling routine in order to weight pharmacophore selection toward filling the diversity voids in previously constructed libraries.

$$\text{Conscore} = \sum_{i=1}^{a} O_i S_i \quad (3)$$

where Conscore = constraint score, O_i = number of times pharmacophore i has been hit for molecules selected from current data set, S_i = score associated with pharmacophore i for the constraining library, a = number of accessible pharmacophores.

$$S_i = [\max(0,(\text{av cov} - Oc_i))]^{\nu} \quad (4)$$

where $\max(0,\text{av cov} - Oc_i)$ = maximum of the values 0 and av cov$-Oc_i$, av cov = the average pharmacophore count across all occupied pharmacophores in constraining library, Oc_i = number of molecules containing pharmacophore i in the constraining library, ν = user defined weight.

$$\text{av cov} = \frac{\sum_{i=1}^{a} \min(Oc_i,\beta)}{Unique_c} \quad (5)$$

where $\min(Oc_i,\beta)$ = minimum of the values Oc_i and β, β = user-defined maximum contribution to av cov by any single pharmacophore, $Unique_c$ = number of pharmacophore bins occupied in constraining library.

(v) To allow the user to weight the score against promiscuous molecules (structures which exhibit a large number of pharmacophores), the total number of pharmacophores present in all currently selected molecules (*Totpharm*) is also included in our "energy" function.

(vi) Finally, the total number of scoring molecules (S) in the selected set is also included. This is used to weight the selected data set to include more molecules that pass user defined bounds of acceptability (for example based on maximum flexibility or pharmacophore promiscuity). All these features are combined to create our overall scoring function.

$$energy = (Unique^{w} \times \text{Conscore} \times \text{Partscore}_{pp}{}^{x} \times \text{Partscore}_{pa}{}^{x} \times \text{Partscore}_{ha}{}^{x} \times S)/(Totpharm^{y} \times \text{Flex}^{z} \times n) \quad (6)$$

where w, x, y, z = user defined weights.

Experimental Section

Two data sets were used to study the behaviour of HARPick, both in general terms and with respect to its nearest relative, Chem-Diverse:

(i) 20168 molecules taken from the Standard Drug File (SDF).[26] All the molecules chosen had between 15 and 51 heavy atoms (excluding halogens), a simplistic criterion for choosing molecules with "druglike" size. Note that we employ this technique as a general screen for molecule size. In the case of the SDF, it is less relevant, although many of the smaller molecules removed tend to be of less interest (e.g. antiseptics), while most of the larger structures are peptides. All these structures were converted to 3D using Concord.[27]

(ii) A simple hypothetical combinatorial library comprising two components undergoing amide bond formation (Figure 3). The acids and amino acids for the library were selected from the available chemicals database (ACD).[28] The reagents selected were constrained to be of between 8 and 25 heavy atoms (excluding halogens) so that the products matched the size of the molecules selected from the SDF (again a "druglike" size). Molecules with a heteroatom ratio (ratio of the number of heteroatoms, excluding halogens, versus the total number of heavy atoms in the molecule) outside the 0.1−0.5 range (the range in which >90% of the molecules in the SDF fall) were then removed, as were structures containing undesirable (e.g. toxic and reactive) groups.[15]

It should be noted at this point that, although we would prefer to profile the complete virtual library, the total number of reagents available to many combinatorial libraries (including the one used here) make such an approach prohibitive. We must therefore filter the reagent pools down to a size we can deal with in product space (<100 000 products) using simpler and more rapidly calculable descriptors. To this end, reagents were clustered using 2D Daylight[29] fingerprints and Wooton spheres[30] at a similarity level of 0.775. This methodology is designed to provide a rapid means for discarding molecules with similar 2D structures.

All the molecules passing these tests were converted into 3D using Concord. Six acids were found to cause problems upon conversion to 3D and these were also removed. This left 67 amino acids and 505 acids, giving a total library size of 33 835 products. Both the SDF and hypothetical libraries were then profiled using Chem-X/Chem-Diverse (July96 version) software, employing our own PCL script. The resulting molecular pharmacophore profiles, heavy atom count and flexibility values were stored on disk. The SDF required around 1 day to be profiled on an SGI 195 MHz R10000 (upon which all calculations were undertaken). Around 6 days were required to profile the hypothetical library structures. Five experiments were undertaken using these data sets to analyze the performance of HARPick.

(1) A full pharmacophore profile of the SDF data set was calculated to show the pharmacophore distribution across a typical molecular collection.

(2), (3) Chem-Diverse and HARPick were applied to the task of compound subselection from the SDF, under various conditions.

(4) Calculations employing the hypothetical library were executed, on this occasion to highlight the performance of HARPick when selecting compounds from multicomponent data sets.

(5) Subselections were made from the hypothetical library with reference to the full SDF profile to illustrate the ability of HARPick to choose constrained libraries.

Results

(1) The full SDF profile was studied to illustrate the nonbinary nature of pharmacophore profiles. Structure

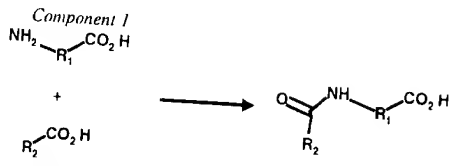

Figure 3. Two component hypothetical library used in the HARPick experiments, comprising "amino acid" (component 1) and acid (component 2) reagents.

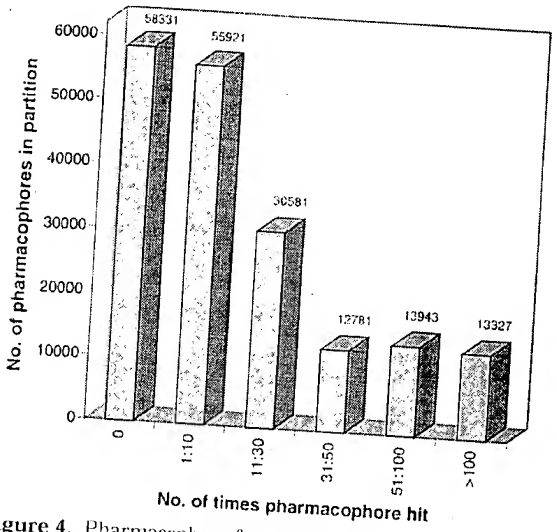

Figure 4. Pharmacophore frequency distribution histogram for 20 169 molecules taken from the SDF (study 1). Total number of pharmacophore in profile = 4 797 745. Number of geometrically accessible pharmacophores = 184 884. Number of different pharmacophore triplets found in library = 126 553. Thus over 68% (126553/184884) of the accessible pharmacophores are present in the library.

pharmacophores with a perimeter ratio of less 0.7 relative to the largest pharmacophore perimeter found in the same molecule were removed. The resulting histogram of pharmacophore distributions is illustrated in Figure 4.

(2) Diverse subset selections of the SDF were made using Chem-Diverse and HARPick to illustrate their diversity profiling characteristics.

(i) Chem-Diverse was used to select a subset of the library based on pharmacophore diversity. The molecules were ordered according to heavy atom count (smallest to largest) to try to force the selection of smaller pharmacophores from smaller structures. The maximum pharmacophore overlap percentage permitted between each keyed molecule and the total library key was set to 60%. All molecules in the library were processed, with any passing the selection criterion being added to the selected subset. This calculation resulted in the selection of 372 molecules.

(ii) Three HARPick runs were then undertaken on the same SDF data set, with the program set to select an identical set size to Chem-Diverse (372) using different diversity criteria: (a) Maximize the internal pharmacophore diversity only. The following values and weights were applied to the diversity function (see eqs 3 and 6): Conscore = 1, $w = 1$, $x = 0$, $y = 0$, $z = 0$. (b) Maximize the internal pharmacophore diversity while maximizing the shape partition scores. Diversity function values and weights applied: Conscore = 1, $w = 1$, $x = 1$, $y =$

Table 1. Results from Study 2

calculation[a]	no. of unique pharmacophores[b]	total no. of pharmacophores	property partition function scores			no. of calculable conformers	run parameters *Conscore, w, x, y, z*
			ha	*pp*	*pa*		
Chem-Diverse 2(i)	49829	68987	0.33	0.66	0.43	2.2×10^8	not applicable
	{71096[c]}						
HARPIck 2(ii)(a)	105222	419870	0.97	0.69	0.70	4.0×10^8	1, 1, 0, 0, 0
HARPick 2(ii)(b)	61913	93977	1.00	0.90	0.85	6.4×10^8	1, 1, 1, 0.75, 0
HARPick 2(ii)(c)	70656	137801	0.90	0.79	0.54	2.4×10^6	1, 2, 0.5, 0.75, 0.33
	{87566[c]}						
random 2(iii)	39625	80556	0.54	0.47	0.34	1.1×10^8	not applicable

[a] Time taken for Chem-Diverse calculation ~22 h. All primary HARPick calculations ran for between 8 and 22 min (10000–25000 iterations at a speed of ~40 iterations per second). [b] Number of unique pharmacophores present that pass the 0.7 perimeter ratio test. The Chem-Diverse results have been converted to reflect this. [c] Unique pharmacophore score in unrestricted space (no perimeter filter applied). This was the pharmacophore space used during the Chem-Diverse selection procedure.

0.75, $z = 0$. (c) Maximize the internal pharmacophore diversity while maximizing the shape partition scores and minimizing the flexibility. Diversity function values and weights applied: Conscore = 1, $w = 2.0$, $x = 0.5$, $y = 0.75$, $z = 0.33$.

(iii) To provide a reference set, the average data from three random selections of 372 molecules were also collated.

Note: For all the above HARPick calculations, structure pharmacophores with a perimeter ratio of less than 0.7 relative to the largest pharmacophore perimeter found in the same molecule were removed. The resulting diversity profiling data are shown in Table 1 and Figure 5.

(3) All promiscuous molecules (structures containing >1500 pharmacophores or >10 000 calculable conformers) were removed from the SDF set. The remaining 15 716 structures were again profiled using Chem-Diverse and HARPick using different search conditions.

(i) For a second time, Chem-Diverse was used to select a subset based on pharmacophore diversity. For this calculation molecules were ordered randomly, as recommended by Chemical Design for large data sets. The maximum pharmacophore overlap percentage permitted between each keyed molecule and the total library key was once more set to 60%. In addition, a pharmacophore area to heavy atom count ratios of 0.4 was enforced. The calculation was set to terminate after the selection of 400 molecules, which Chem-Diverse achieved after processing 4500 structures.

(ii) Five HARPick runs were then undertaken on the same SDF data set, with the program set to select an identical set size to Chem-Diverse (400) using various diversity criteria: (a) Maximize the internal pharmacophore diversity only. Diversity function values and weights applied: Conscore = 1, $w = 1$, $x = 0$, $y = 0$, $z = 0$. (b) Maximize the internal pharmacophore diversity while maximizing the shape partition scores. Diversity function values and weights applied: Conscore = 1, $w = 1.5$, $x = 1$, $y = 0.75$, $z = 0$. (c) Maximize the internal pharmacophore diversity while maximizing the shape partition scores and minimizing the flexibility. Diversity function values and weights applied: Conscore = 1, $w = 2.0$, $x = 0.5$, $y = 0.75$, $z = 0.33$. (d) Maximize the internal pharmacophore diversity while trying to minimize the pharmacophore promiscuity of the selected molecules. Diversity function values and weights applied: Conscore = 1, $w = 1.75$, $x = 0$, $y = 1.0$, $z = 0$. (e) Try to balance all the nonconstraint elements of the diversity function. Diversity function values and weights applied: Conscore = 1, $w = 1.75$, $x = 0.25$, $y = 1.0$, $z = 0.33$.

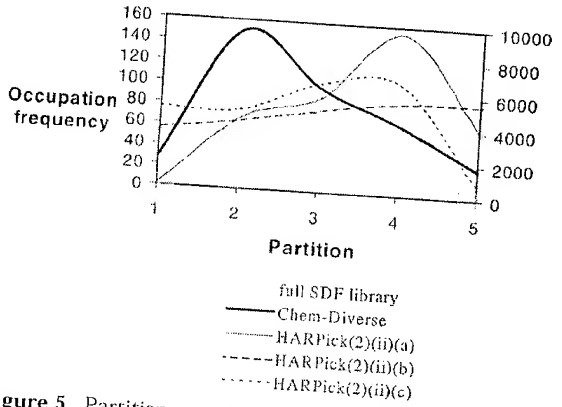

Figure 5. Partition occupation distributions for largest molecular pharmacophore perimeters (*pp*—see Equations 1 and 6) from study 2. Left-hand *y*-axis and lines show distributions for selected subsets of the SDF, while right-hand *y*-axis and columns show distribution across the whole library.

(iii) To provide a reference set, the average data from three random selections of 400 molecules were also collated.

Note: For all the above HARPick calculations, structure pharmacophores with a perimeter ratio of less than 0.7 relative to the largest pharmacophore perimeter found in the same molecule were removed. The resulting diversity profiling data are given in Table 2.

(4) The fourth investigation was undertaken to illustrate the differences between HARPick and Chem-Diverse in multiple component product profiling. On this occasion, primary comparisons were made in terms of reagent selection control. The data set chosen for analysis was a subset of the hypothetical library shown in Figure 3. All 67 reagents from component 1 and the first 19 reagents from component 2 were selected (the full set was not used since it would require nearly 6 days CPU to be profiled in Chem-Diverse). Two experiments were executed using this data set.

(i) Chem-Diverse was used to select a subset of the library based on pharmacophore diversity. The molecules were ordered using the Chem-X "sample" command, which tries to ensure maximal difference in F-group keys (Chemical Design 2D fingerprints[5]) at the start of the list. The maximum pharmacophore overlap percentage permitted between each keyed molecule and the total library key was set to 95% (higher than in the previous studies as the molecules are much more closely

Table 2. Results from Study 3

calculation[a]	no. of unique pharmacophores[b]	total no. of pharmacophores	property partition function scores			no. of calculable conformers	run parameters *Conscore, w, x, y, z*
			ha	*pp*	*pa*		
Chem-Diverse 3(i)	37811	58391	0.8	0.67	0.47	3.7×10^5	not applicable
HARPick 3(ii)(a)	73999	237656	0.82	0.55	0.57	1.1×10^6	1, 1, 0, 0, 0
HARPick 3(ii)(b)	55677	99180	0.89	0.92	0.60	4.9×10^5	1, 1.5, 1, 0.75, 0
HARPick[c] 3(ii)(c)	55156	100997	0.80	0.86	0.57	3.1×10^5	1, 2, 0.5, 0.75, 0.33
	{61207}	{109371}	{0.88}	{0.92}	{0.61}	$\{3.1 \times 10^5\}$	
HARPick 3(ii)(d)	50811	69727	0.46	0.55	0.36	4.1×10^5	1, 1.75, 0, 1, 0
HARPick 3(ii)(e)	49994	78191	0.71	0.71	0.55	3.1×10^5	1, 1.75, 0.25, 1, 0.33
random 3(iii)	26992	56102	0.45	0.37	0.28	3.1×10^5	not applicable

[a] Time taken for Chem-Diverse calculation ~4 h. All primary HARPick calculations ran for between 5 and 20 min (10000–25000 iterations at a speed of ~50 iterations per second). [b] Number of unique pharmacophores present that pass the 0.7 perimeter ratio test. The Chem-Diverse results have been converted to reflect this. [c] Long simulated annealing run with same diversity function shown in brackets. This job required ~3 CPU hours to execute 500 000 iterations.

Table 3. Results from Study 4

calculation	no. of unique pharmacophores	total no. of pharmacophores	no. of reagents selected from component 1	no. of reagents selected from component 2	no. of calculable conformers	run parameters *Conscore, w, x, y, z*
Chem-Diverse 4(i)	55281	290131	23	13	2.2×10^7	
HARPick 4(ii)	50951	235944	10	5	1.8×10^7	1, 1, 0, 0, 0

Table 4. Results from Study 5

calculation[a]	no. of unique pharmacophores	total no. of pharmacophores	*Conscore* / *Unique*[b]	no. of pharmacophores found in scoring bins	run parameters *Conscore, w, x, y, z*
HARPick 5(i)(a)	78567	806539	3.4	99696 (12%)	1, 1, 0, 0.5, 0
HARPick 5(i)(b)	79125	1079998	7.9	203061 (19%)	0, 1, 0, 1, 0
random 5(ii)	51791	1033772	2.3	51837 (5%)	$\nu = 1, \beta = 10$ not applicable

[a] HARPick calculations ran for around 30 min (~20000–25000 iterations at speeds of around 11[5(i)(b)]–16[5(i)(a)] iterations per second). [b] See eqs 3–6.

related than in the SDF). All molecules in the library were processed, with any passing the selection criterion being added to the chosen subset. This calculation resulted in the selection of 50 molecules.

(ii) HARPick was then run, maximizing the internal pharmacophore diversity of 50 compounds. Rather than being allowed to select products in an unconstrained manner (as we must in Chem-Diverse), the program was forced to select only 10 reagents from component 1 and 5 reagents from component 2. The following values and weights were applied to the diversity function: Conscore = 1, $w = 1$, $x = 0$, $y = 0$, $z = 0$. On this occasion no pharmacophore perimeter filters were applied to the molecular pharmacophore descriptors. Results for study (4) are given in Table 3.

(5) The fifth and final study was used to illustrate the ability of HARPick to build constrained libraries. The full hypothetical library of our two component data set (Figure 3) was chosen as the library to be profiled.

(i) Two HARPick runs were undertaken, with the calculation constrained to select 20 reagents from component 1 and 50 from component 2: (a) Maximize internal pharmacophore diversity only. The following values and weights were applied to the diversity function: Conscore = 1, $w = 1$, $x = 0$, $y = 0.5$, $z = 0$. (b) Starting with the selections made in run (5)(i)(a), maximize internal pharmacophore diversity. At the same time, weight the calculation towards pharmacophores which fill relative voids in the SDF library profile shown in Figure 4. The following weights were applied to the diversity function: $\nu = 1$, $w = 1$, $x = 0$, $y = 1$, $z = 0$. The β value applied to the av cov calculation (eq 5) was set to 10, which lead to a maximum scoring pharmacophore occupation level of 7 for the constraint term. The minimizing function (rather than the full simulated annealing search procedure) was employed for this calculation.

(ii) To provide a reference set, the average data from three random selections of 20 reagents from component 1 and 50 from component 2 were collated.

Note: For both the above HARPick calculations, structure pharmacophores with a perimeter ratio of less than 0.7 relative to the largest pharmacophore perimeter found in the same molecule were removed. The resulting diversity profiling data are given in Table 4.

Discussion

Study 1 (Figure 4) clearly illustrates the problems of a binary profile. The SDF library studied contains only 20 168 molecules and a strict perimeter filter was applied, yet over 68% of the total geometrically accessible pharmacophores are present in the data set. It should be noted that a significant number of these accessible pharmacophores might well be considered unsuitable in medicinal chemistry terms (e.g. acid–acid, lipophilic–lipophilic–lipophilic, all three distances > 20 Å etc.), thus the actual occupation level is almost certainly higher. When no perimeter filter is applied, over 15 million pharmacophores are present in the library, which is 3 times the number present in the profile used. If all of these pharmacophores were included, the number of occupied bins would clearly be even greater. It is thus obvious that saturation of a

binary key will quickly become a problem. In addition, the histogram clearly shows the unevenness of the pharmacophore distribution. While over 50 000 pharmacophores are hit between 1 and 10 times, around 40 000 are hit more than 30 times. If we are to take full advantage of pharmacophore keys for constrained library design, it is evident that we will need to exploit this distribution information (as HARPick attempts to, using eq 3). Studies 2 and 3 (Tables 1 and 2) clearly illustrate the many advantages of a customizable diversity function when addressing the profiling problem. Both Chem-Diverse and HARPick are able to considerably improve molecular selection based on pharmacophore count, compared to random selections (2(iii) and 3(iii)). HARPick calculations 2(ii)(a) and 3(ii)(a), which were set to purely maximize pharmacophore diversity, are able to find around twice the number of pharmacophores of the comparable Chem-Diverse runs. As one would expect, however, the molecules chosen are substantially more flexible and promiscuous (as evidenced by the total pharmacophore and calculable conformer counts) and are also not optimally partitioned in our simplified version of shape space. The remaining HARPick calculations in studies 2 and 3 illustrate how we can address these various selection features through simple customization of the diversity scoring function. Calculations 2(ii)(b) and 3(ii)(b) show how the inclusion of the partition function (eq 1) in the diversity score considerably improves the shape property partitioning, while still allowing good pharmacophore diversity. Figure 5 illustrates the effect of including the partition function (eq 1) in the diversity score. Chem-Diverse is seen to broadly follow the trends present in the whole SDF library with respect to perimeter distribution. Study 2(ii)(a), which is only tuned to maximizing internal pharmacophore diversity, shows a preponderance of larger perimeter selections. Studies 2(ii)(b) and 2(ii)(c), however, both of which include a partition weighting in their diversity score, show a significantly more even partitioning of perimeters. Studies 2(ii)(c) and 3(ii)(c) illustrate how the addition of a flexibility minimization function (eq 2) substantially reduces the number of calculable conformers present in the selected data sets. Even when we analyze the unique pharmacophore count in unfiltered space as applied in the Chem-Diverse calculations (Table 1, footnote *c*), the HARPick selection still contains substantially more pharmacophores. HARPick run 3(ii)(d) shows how increasing the weighting for the *Totpharm* denominator term (eq 6) can dramatically improve the ratio of unique pharmacophore occupation to total pharmacophore count (i.e. reduce the pharmacophore promiscuity ratio *Unique*/*Totpharm*). Finally, calculation 3(ii)(e) illustrates that even with the inclusion of all the nonconstraint functions, the resulting HARPick selection is able to outperform the Chem-Diverse data set. Both flexibility and partition scores are improved, over 12 000 extra pharmacophore types are found, and a comparable pharmacophore promiscuity ratio is maintained (64% for HARPick versus 65% for Chem-Diverse). Another Chem-Diverse problem highlighted by study 3 is the fact that only 4500 structures were profiled to achieve the required selection of 400 molecules. The only way to control the size of the selected data set is to force the calculation to terminate after a fixed number of structures have been chosen. This means that, unlike in HARPick, there is no way to sample the whole of available structure space.

Study 4 illustrates the advantage of decoupling the diversity calculation from the selection procedure (Table 3). Chem-Diverse selected 50 products from the available product data set. Rather than choosing the most efficient combination of reagents (15), however, Chem-Diverse selected products containing 36 different constituents. This could be considered as a selection of 61% efficiency, setting 15 as 100% efficient and 69 (the largest number of reagents which could be selected from 50 products of this two component data set) as 0% efficient ($|69 - 36| \times 100/|69 - 15|$). In contrast, it is a simple matter in HARPick to define the requirement that a 50 product selection from a two component reaction contain 10 reagents from component 1 and 5 reagents from component 2. Thus a 100% efficient selection can be made with ease, and in this case the resulting data set is of comparable quality to the significantly less efficient set chosen by Chem-Diverse. This is of substantial importance, since many chemists wish to create multidimensional libraries of 100% efficiency, both on grounds of cost and ease of robot programming. The final study illustrates the advantages of applying nonbinary pharmacophore constraints to diversity profiling calculations. As one would expect, both constrained and unconstrained HARPick runs outperform random selections from the perspective of filling diversity voids. The unconstrained HARPick search (5(i)(a)) does significantly better than random (study 5(ii)) because, although the SDF profile contains many pharmacophores, its relatively small size still leaves substantial holes in pharmacophore space. As a consequence, maximizing pharmacophore spread is bound to increase the constraint score. Nevertheless, as soon as we constrain our selection specifically to fill these voids (calculation 5(i)(b)), considerable improvements are observed. While the internal diversity of the system is maintained, the average constraint score per pharmacophore type is found to more than double, and the number of pharmacophores found in scoring bins as a proportion of the total pharmacophores present increases from 12% to 19%. In Chem-Diverse it is currently only possible to count pharmacophores which have not been hit at all in the constraining libraries as voids (the binary key problem). Also there is no way to couple intra- and interlibrary diversity in the calculation. Combine this with the 6 CPU days required for a single Chem-Diverse calculation on the library, and it is clear that a comparable calculation would be impractical.

The results in study 5 highlights a feature of complex scoring functions. Intuitively, one would expect the number of unique pharmacophores found in study 5 calculations to be greater for the unconstrained search (calculation 5(i)(a)). In fact we find that the constrained search (calculation 5(i)(b)) contains slightly more unique pharmacophores. This is because, when we include the Conscore (eq 3) function, its nonbinary nature allows more pharmacophore space to contribute to the diversity score. As a consequence the relative contribution of the *Totpharm* denominator term is reduced, allowing the presence of more promiscuous molecules. This can be seen by the consequent increase in the total number of

pharmacophores seen in Table 4. When the *Totpharm* weight *y* was left at 0.5 for study 5(i)(b), more than 1.25 $\times$ 10^6 pharmacophores were present in the resultant chosen set (data not shown). These results illustrate the relationships that can be formed between terms in a complex scoring function. This emphasizes the need for careful setting of the customizable weights to ensure that optimal selections are obtained. An interesting feature of the HARPick calculations is their speed. Nearly all runs were completed in under 30 min. A direct comparison with a single Chem-Diverse run is difficult, as the profiling calculation required to generate the data for HARPick essentially takes as long as a single Chem-Diverse study on the same data set. Of extreme importance, however, is the fact that a single run rarely suffices when profiling a given library. Problems such as reagent cost, chemists' dislike of certain reagent types, and a suboptimal balance of physicochemical properties in initial selections can all lead to the requirement for multiple profiling calculations. If one considers this in the context of the hypothetical library studied here, the implications are clear. Since a single Chem-Diverse profiling calculation requires nearly 6 days to run, multiple profiles of this sort become completely untenable. Conversely, with HARPick, once the pharmacophore data have been calculated and stored, actual profiling calculations can be undertaken rapidly, rendering iterative profiling a simple process.

While the HARPick runs studied all converged to excellent solutions, finding the global minimum can be somewhat problematic. The annealing schedule is not always finding the global minimum, and in some cases longer runs can improve results; we have also found that it is possible to obtain good results in a short time span. This reflects the stochastic nature of the annealing protocol. We have experimented with several schedules but could not find a universal recipe. We recommend using a quick cooling schedule to locate an approximate minimum, followed by a review of the results before trying a longer schedule. We also note that, given that our diversity measures are not the absolute truth, a near-global minimum may be just as valid a solution as a global minimum of comparable depth. Repeated calculations for study 3 (data not shown) using identical annealing calculations led to selections with an average of 26% of molecules in common. Repeated calculations of the longer (500 000 iteration 3(ii)(c)) run (data not shown) lead to selections with 51% of molecules in common. Achieving a 100% identical selection for this study is difficult, however, as no attempt has been made to remove very similar molecules from the system. It would thus be simple for HARPick to substitute near-identical structures into the system and achieve the same quality profile (as we observed from the 3(ii)(c) run statistics (data not shown), which were found to be near-indistinguishable). For calculations involving pre-clustered (and hence nonidentical) reagents, as in study 4, repeated runs were able to converge to the same data set selections. As with many such stochastic problems, however, the ability to find the global minimum will be tied to many factors, the primary ones here being the size and constitution of the data set. Nonetheless, these studies show that it is possible to produce good results over relatively short time scales. It should be emphasized that the studies we have undertaken are on the same scale as would be envisaged in practical profiling calculations. Indeed, the hypothetical library used contains unusually promiscuous molecules ($\sim$35 $\times$ 10^6 pharmacophores across 33 835 molecules with the pharmacophore perimeter filter being applied), with the average structure containing on average 4 times the number of pharmacophores present in a molecule of the SDF.

The computational complexity of the algorithm is hard to define because of its heuristic nature. 90% of cpu time is taken up in pharmacophore evaluation, which, to a first approximation, scales linearly with the number of pharmacophores in the set. Use of a less compute intensive primary descriptor would lead to a significant increase in program speed. Furthermore, although collating the initial profiling data for HARPick across the full product data set can be time consuming (6 days CPU in the case of the hypothetical library studied here), the profiling procedure lends itself easily to parallel calculation. It is thus a simple task to spread the profiling calculation across all available CPUs, dramatically reducing the time required to collate the pharmacophore data. This phase scales as the number of products in the set, linked to granularity of the pharmacophoric analysis. In the annealing phase, the number of products evaluated crudely varies as the square of the number of reagents chosen. The rate of convergence depends on the cooling schedule and the redundancy in the set (i.e. the degree of separation of the global minimum from other minima). Of these factors, our guess is that the controlling one will be the number of reagents, so that the algorithm has an approximately quadratic dependence. However, we note that the RAM requirements will probably prove a more demanding constraint than the algorithmic complexity. Currently, the program stores all the pharmacophores present in a product data set in RAM for easy access. To store the 35 million pharmacophores of the combinatorial library studied here requires around 140 Mbytes of RAM. One could envisage a modification of HARPick, however, where only the current selected set of products are held in RAM. All remaining pharmacophore profiles would be stored on disk and accessed as required by the optimization algorithm reagent selections. With this program structure, only around 5 Mbytes would have been required to store the 1000 selected products from study (5). The performance implications are not clear, however, and from our own perspective, easy access to cheap RAM has made such a modification a low priority.

The exact nature and type of stochastic optimization algorithm that would give the best results needs further investigation. While the simulated annealing protocol used above was found to work well, the random nature of reagent selection suggests the efficiency of the procedure might be improved upon. Intuitively, one would imagine that the selection of reagents possessing properties unique to a library would generally lead to an improvement in the fitness of the chosen data set. An algorithm which is able to retain a "memory" of previous mutation quality[31] may well be found to converge more rapidly to a preferred solution.

The primary feature emphasized by the above calculations is control. The application of a stochastic optimization algorithm allows us command over both

the number of reagents chosen for each component, and the nature of the diversity function used to select them. The potential flexibility of such an approach is clear. In principle, any descriptor could be applied to the scoring functions. One could envisage maximizing functions (e.g. 3D pharmacophore or 2D fingerprint coverage/reagent supplier quality), minimizing functions (e.g. cost per reagent), partition functions (general shape/log P) and bounding functions (zero score products with properties outside bounds, e.g. minimum/maximum log P). In principle, a totally customizable scoring function could be devised, with the user able to choose which properties are included in the scoring routine, and the functions used on them. With careful application of user weightings for each component function, the result would be a totally flexible profiling paradigm. This is currently an area of active research.

Conclusions

The objective of this research was to tackle the problem of combinatorial library design, in a manner that answered the needs of our medicinal chemists. The methodologies described above overcome many of the inherent deficiencies present in first generation diversity tools. The techniques allow efficient storage of pharmacophore descriptors, explicit reagent selection during product-based diversity analysis, easy incorporation of alternative user-defined parameters, plus more extensive profiling tools to allow library designs constrained by already synthesized product databases. All of these features provide the basis for a highly versatile profiling paradigm which should prove extremely useful in library design.

Acknowledgment. Thanks to all our colleagues at Rhône-Poulenc Rorer for their helpful comments and suggestions during the creation of this manuscript, particularly David Clark, Paul Bamborough, Jon Mason, Stephen Pickett, and Chris Newton.

References

(1) Gordon, E. M.; Gallop, M. A.; Patel, D. V. Strategy and Tactics in Combinatorial Organic Synthesis. Applications to Drug Discovery. *Acc. Chem. Res.* **1996**, *29*, 144−154.

(2) Willett, P. Computational Tools for the Analysis of Molecular Diversity. In *A Practical Guide to Combinatorial Chemistry*; Czarnik, A.; Hobbs-Dewitt, S., Eds.; ACS Books: Washington, D.C., 1997; in press.

(3) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.

(4) Brown, R. D.; Martin, Y. C. Use of Structure Activity Data to Compare Structure-Based Clustering and Descriptors for Use in Compound Selection. *J. Chem. Int. Comput. Sci.* **1996**, *36*, 572−584.

(5) Chem-Diverse, developed and distributed as part of the Chem-X suite of modelling software by Chemical Design Ltd., Roundway House, Cromwell Business Park, Chipping Norton, OXON, UK.

(6) Davies, E. K.; Briant, C. *Combinatorial Chemistry Library Design Using Pharmacophore Diversity*. Accessible through URL http://www.awod.com/netsci/Science/Combichem/feature05.html.

(7) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(8) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*, Vol. 7; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995; pp 1−66.

(9) Cramer, R. D.; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443−485.

(10) Marshall, G. R. Binding Site Modelling of Unknown Receptors. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 80−116.

(11) Good, A. C.; Mason, J. S. Computational Screening of 3D Databases. In *Reviews in Computational Chemistry*, Vol. 7; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1995; pp 67−117.

(12) Wang, S. M.; Milne, G. W. A.; Yan, X. J.; Posey, I. J.; Nicklaus, M. C.; Graham, L.; Rice, W. G. Discovery of Non-Peptide HIV-1 Protease Inhibitors by Pharmacophore Searching. *J. Med. Chem.* **1996**, *39*, 2047−2054.

(13) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Int. Comput. Sci.* **1997**, *37*, 731−740.

(14) Pearlman, R. S. *Novel Software Tools for Addressing Molecular Diversity*. Accessible through URL http://www.awod.com/nersci/Science/Combichem/feature08.html.

(15) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analyses of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach. *J. Chem. Int. Comput. Sci.* **1997**, *37*, 599−614.

(16) Good, A. C.; Kuntz, I. D. Investigating the Extension of Pair-Wise Distance Pharmacophore Measures to Triplet-Based Descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373−379.

(17) Pickett, S. D.; Mason, J. S.; Mclay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore Derived Queries (PDQ). *J. Chem. Int. Comput. Sci.* **1996**, *36*, 1214−1233.

(18) Willett, P.; Winterman, V. Implementation of Nonhierarchical Cluster Analysis Methods In Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Int. Comput. Sci.* **1986**, *26*, 109−118.

(19) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(20) Van Laarhoven, P. J. M.; Aarts, E. H. L. *Simulated Annealing: Theory and Applications*; Reidel: Dordrecht, 1987.

(21) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and Visualization of Molecular Diversity of Combinatorial Libraries. *Molecular Diversity* **1996**, *2*, 64−74.

(22) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(23) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Int. Comput. Sci.* **1995**, *35*, 310−320.

(24) Agrafiotis, D. K.; Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(25) Mason, J. S. Experiences with Searching for Molecular Similarity in Flexible 3D Databases. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: Glasgow, 1995; pp 138−162.

(26) Standard Drug File (now known as the World Drug Index), Derwent Publications Ltd., 14 Great Queen Street, London, WC2B SDF, UK.

(27) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. University of Texas, Austin. CONCORD, *A Program for the Rapid Generation of High Quality Approximate 3D Molecular structures*. Distributed by Tripos Associates, 1699 S Hanley, Suite 303, St Louis, MO, USA. Version 3.24 was used for these experiments.

(28) Available Chemicals Database, developed and distributed by Molecular Design Ltd., San Leandro, CA.

(29) Daylight Fingerprint Toolkit. Daylight Chemical Information Systems Inc. 27401 Los Altos, Suite 370, Mission Viejo, CA 92691.

(30) Wooton, R.; Cranfield, R.; Sheppey, G. C.; Goodford, P. J. Physicochemical Activity Relationships in Practice. 2. Rational Selection of Benzenoid Substituents. *J. Med. Chem.* **1975**, *18*, 607−613.

(31) Clark, D. E.; Westhead, D. R. Evolutionary Algorithms in Computer-Aided Molecular Design *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.

JM9704031